Machine Intelligence Project

Group members: Bora Berk 03709300; Serdar Doruk Şenbayrak 03696074; Georgina Joy 03730465; Umut Ekin Gezer 03716498

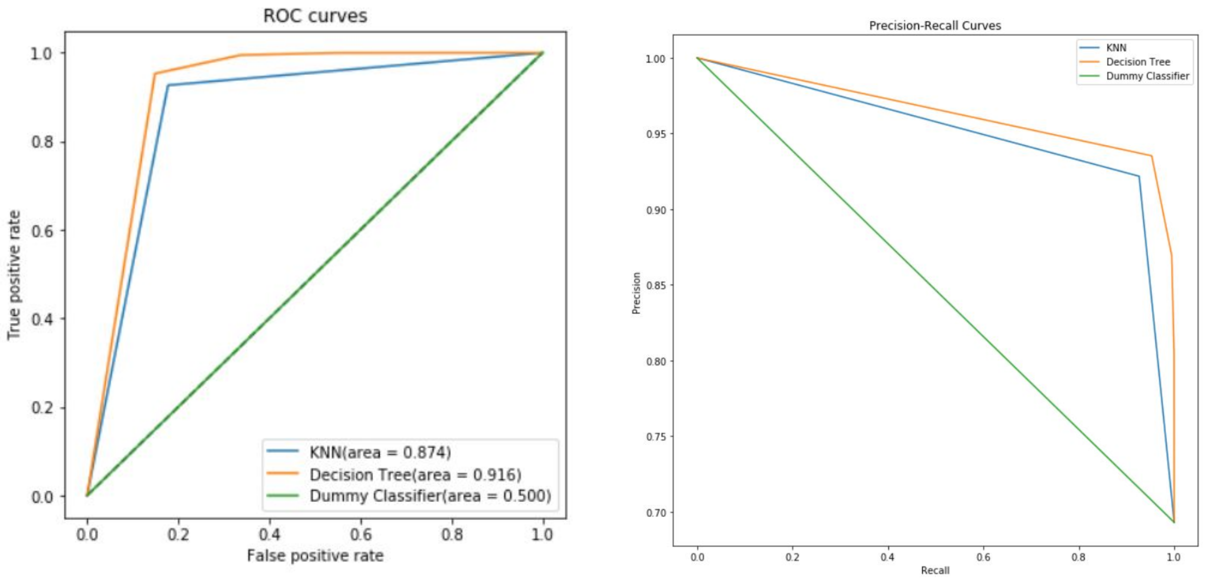
Data pre-processing

This task is based on 2015 traffic stop data sourced from the North Carolina Police Department. As might be expected, the raw data is unsuitable for direct use in machine learning. Therefore, the immediate course of action must be to clean the data and label it correctly. Some variables in the data were considered extraneous and therefore removed from the dataset. This includes the state, as it is data specifically from North Carolina, the column containing what is essentially a copy of information about the driver's race, and the columns containing the stop date and the officer ID, as the goal is to look at the police department as a whole, and therefore a level of consistency in the decisions taken must be assumed.

It's worth noting that both race and search type were encoded with one-hot encoding as opposed to label encoding. Both of these variables were not ordinal, and thus without clear hierarchy. Both also were composed of few enough options that adding the extra columns did not cause too much strain for the model. Furthermore, for the search type data, some entries included multiple options at once, which could only be dealt with by one-hot encoding, and for the race data, this makes certain that there is no value attached to certain races.

Building the models

It was then necessary to decide which algorithms to use for this project, and k-nearest neighbours and decision tree were chosen. This was largely for their utility in this context and because they are able to be manipulated in ways necessary for this set of tasks. A dummy classifier was also created, which always predicts no arrest, to use as a point of comparison.



(a) The ROC curve for all three models.

(b) The precision-recall curve for all three models.

Figure 1: Graphs telling us how good our models are.

As can be seen in Figure 1a, both the kNN and the decision tree model perform significantly better than the dummy classifier, with the decision tree model performing best, as it has the largest area under the curve.

The same is true of the precision-recall curves, in that both of the models are, as expected, a great deal better than the dummy classifier, and again the decision tree model performs better than the kNN model. A high precision means a low rate of false positives and high recall means a low rate of false negatives.

Model	Recall	Precision	F1
kNN	0.927	0.922	0.924
Decision tree	0.975	0.923	0.948
Dummy classifier	0.000	1.000	0.693

Table 1: Metrics to measure the success of our models.

Both precision and recall are important in a functioning model, and at the ends of the curve, a slight improvement in one can lead to a massive fall in the other, leading to a worse outcome.

In order to balance precision and recall, it is useful to calculate the F1 score. This is calculated using these attributes, to balance the demands of each and find the optimal - or fairest, as most would understand it - model, and as can be seen in Table 1, the decision tree model has the highest F1. This is hardly a surprise, given that it has the largest area under the precision-recall curve as well.

These metrics and tests are all used to measure the performance of machine learning algorithms and thus determine their utility in specific contexts. It is clear from Figure 1 and Table 1 that in this context, the decision tree classifier is more effective.

Examining sensitive characteristics

Having created functioning classifiers and evaluated their performance in terms of functionality, it is necessary also to evaluate their performance in terms of fairness. Gender, race, and age were identified as areas of interest, as these are often sensitive characteristics; sensitive characteristics here meaning variables that can unduly influence the data and reproduce real world inequalities and unfairness. The decision tree classifier was used, as it is clearly a more successful method for this data, as can be seen in both Table 1 and Figure 1.

Gender

To begin with, gender was examined, as it only has the two groups. Ideally, the independence calculated for the two groups ought to be the same, as this would indicate that the outcome is statistically independent of the two groups. In this case, as can be seen in Table 2, the value for women is 0.412 and for men, 0.737, giving a difference of 0.325, which is quite high and means that the gender makes a significant difference in outcome.

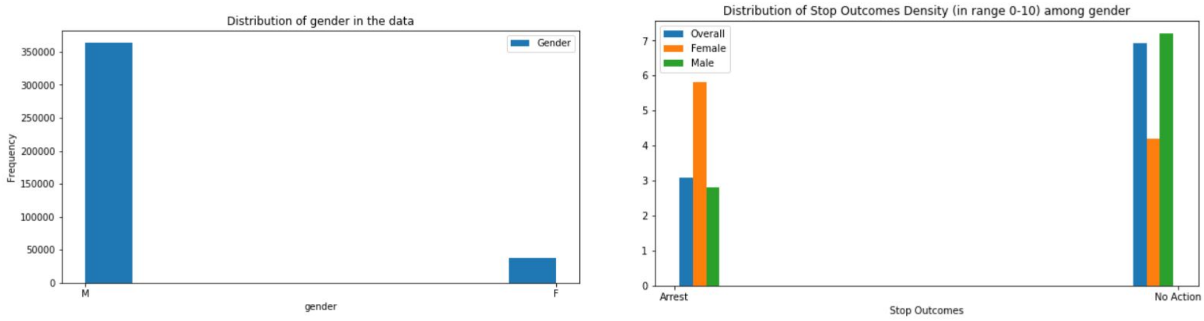
Gender	Independence	Separation (1)	Separation (2)	Sufficiency (1)	Sufficiency (2)
Men	0.737	0.934	0.948	0.983	0.180
Women	0.412	0.751	0.882	0.851	0.203
Difference	0.325	0.183	0.066	0.132	0.023

Table 2: Fairness metric results for gender.

Separation allows correlation between the score and the sensitive attribute, as long as it is justified by the target variable. More simply, it looks at equality of error over groups, and again, the value calculated should be the same across the two groups. Two measures of separation are calculated; one using true positives and the other using true neutrals. For the positive separation value, there is a difference of 0.183, and for the neutrals, a difference of 0.066. Both of these are smaller differences than the independence, which is reassuring.

Sufficiency can also be thought of as calibration - it is checking that the threshold of eligibility is the same across groups. If, in this instance, a man and a woman are in otherwise identical situations, that they would receive the same outcome. Again, there are two measures of sufficiency to be calculated, with one based on true positives and the other on false positives. As before, these values should be as similar as possible across the groups. The value based on true positives returns a difference of 0.132 between the groups, and false positives gives 0.023, which is much better than independence.

Just looking at the data, it appears that the North Carolina police stop far fewer women, and arrest more of those that they do stop, as we can see in Figure 2. Occam's Razor suggests that they are more



(a) Distribution of gender in the data.

(b) Distribution of stop outcome density by gender.

Figure 2: The breakdown of gender in the data.

reluctant to stop women, and thus require a higher level of certainty in their minds that it is necessary to do so before they will stop a woman, leading to very unbalanced data. The model supports this hypothesis, showing that women are disproportionately likely to be arrested, though this provides no information about how often women are likely to be stopped.

Ethnicity

When looking at gender, it is very easy to compare fairness metrics, as there are only two genders measured in this data, and so it is easy to work out the difference in value. With race, this is slightly more complex.

Race	Independence	Separation (1)	Separation (2)	Sufficiency (1)	Sufficiency (2)
Afro American	0.692	0.928	0.909	0.958	0.151
Hispanic	0.448	0.823	0.926	0.900	0.134
White	0.746	0.954	0.871	0.956	0.149
Asian	0.606	0.954	0.825	0.894	0.079
Other	0.611	0.857	0.871	0.913	0.205
Average difference	0.136	0.072	0.048	0.037	0.054

Table 3: Fairness metric results for ethnicity.

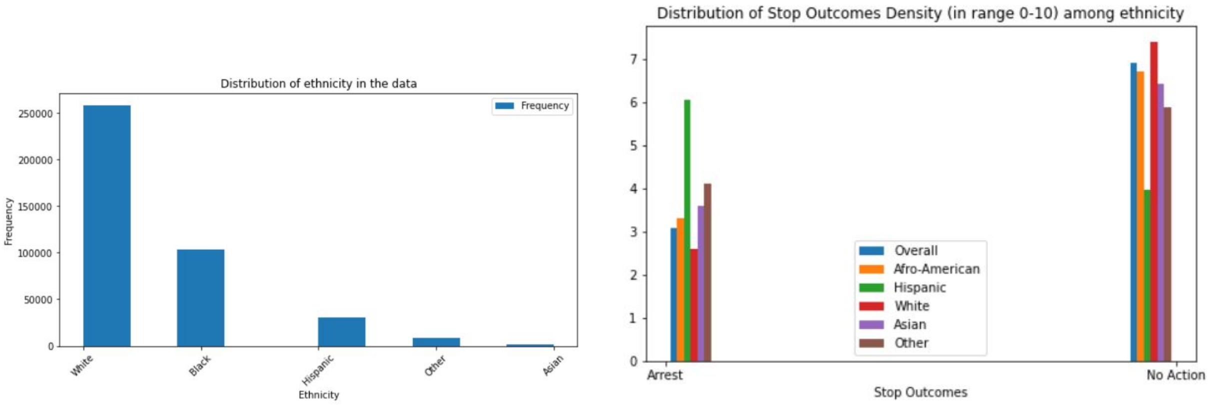
As shown in Table 3, race is actually less sensitive than gender, by nearly every single metric. The only exception is sufficiency measured by false positives, which is slightly higher in race than gender. This might well be explained by the fact that while the data is quite imbalanced by race, it is still less imbalanced than it is by gender. This can be seen in Figure 3, where we can see that the distribution is slightly less stark, despite a clear discrepancy in outcome, particularly between White and Hispanic people.

Age

Age group	Independence	Separation (1)	Separation (2)	Sufficiency (1)	Sufficiency (2)
Young	0.375	0.749	0.918	0.846	0.141
Middle Aged	0.780	0.955	0.885	0.967	0.154
Old	0.884	0.983	0.605	0.950	0.175
Average difference	0.339	0.156	0.209	0.081	0.023

Table 4: Fairness metric results for age.

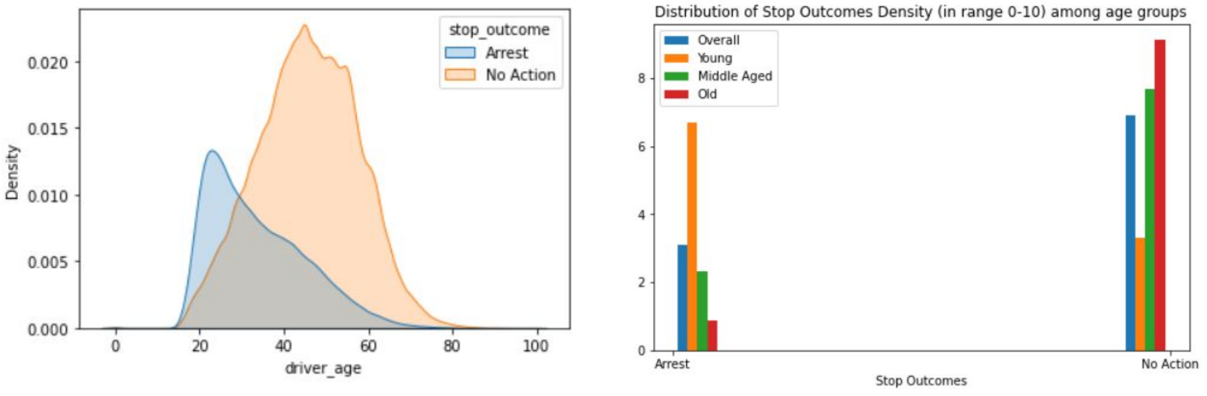
In order to better understand how fair the model is with regard to age, the data was sorted into bins. Figure 5 shows that the age distribution is rather more even than either gender or ethnicity were, but that there is nonetheless still an imbalance in this data. Figure 4a shows the stop outcome density before



(a) Distribution of ethnicity in the data.

(b) Distribution of stop outcome density by ethnicity.

Figure 3: The breakdown of ethnicity in the data.



(a) Distribution of stop outcome density by age.

(b) Distribution of stop outcome density by age, in bins.

Figure 4: Stop outcome density by age in the data.

binning the data, with Figure 4b giving the results after the data has been sorted into these bins that we can use to measure fairness.

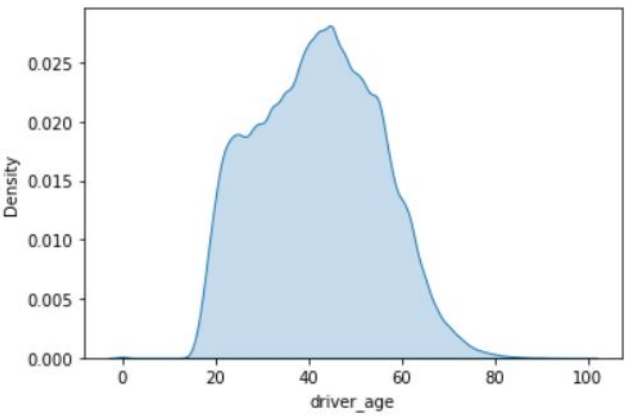


Figure 5: The distribution of age in the data.

In the USA, it is permitted that 14 year olds can drive as a learner, presumably with an adult in the car. Any ages under 14 were therefore ignored - though it is unlikely that there were data points in these region - and the rest sorted into three categories. Those between the ages of 14 and 29 inclusive are designated young, middle aged (for the purposes of this data) is for those between the ages of 30 and 59 inclusive, and anyone 60 or older is considered old.

Figure 4 shows how youth is far more likely to result in an arrest than old age, and even middle age, though to a slightly lesser extent.

The fairness metrics for age are actually rather similar to those for gender, meaning that age is also a very sensitive characteristic. The difference in independence is similarly high, and the difference in separation is higher, when taking both measures of it into account, though the sufficiency for age is a little better than ethnicity.

Model	Independence	Separation (1)	Separation (2)	Sufficiency (1)	Sufficiency (2)
kNN unbiased	0.712	0.923	0.874	0.927	0.178
kNN biased	0.697	0.922	0.832	0.948	0.178
Decision tree unbiased	0.727	0.926	0.927	0.971	0.174
Decision tree biased	0.732	0.923	0.935	0.975	0.184

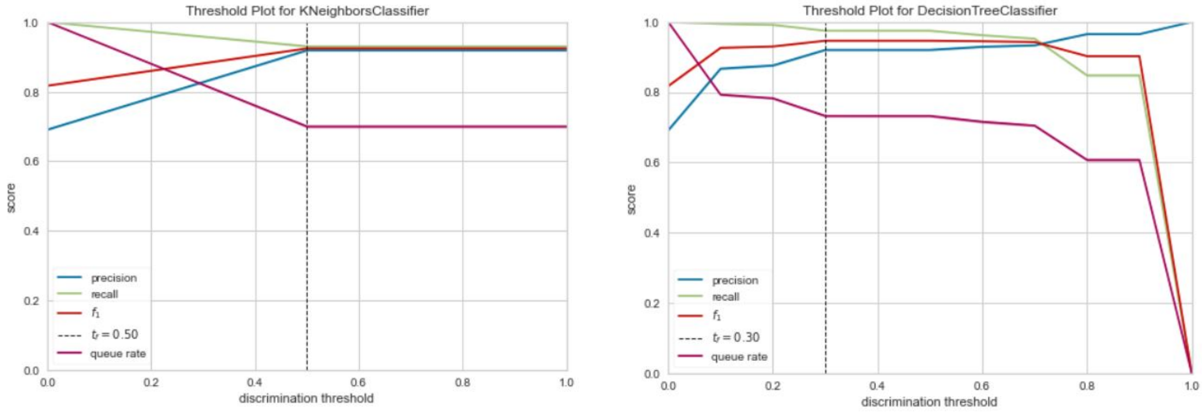
Table 5: Fairness metric results for data sets both with and without sensitive characteristics.

Dropping sensitive characteristics

As age, gender and ethnicity have all proved to be sensitive characteristics, as might reasonably be expected, they were at this point all removed from the data set to investigate whether this improves the fairness of the models.

Compared to the unbiased models, once sensitive characteristics are removed, all the fairness metrics are improved for the k-neighbours classifier, albeit by amounts of the order of 10×10^{-3}. This is less true for the decision tree model, which improves on one measure of sufficiency, but otherwise deteriorates by amounts of the order 10×10^{-4}. These values are shown in Table 5 This was unexpected, as the sensitive characteristics were examined in the context of this classifier, and clearly showed an impact on the fairness of the model. It is possible that this can be explained by the way that the kNN classifier uses all attributes simultaneously, creating a multi-dimensional space in which to do so, the decision tree looks at each attribute one by one and makes its decisions thus. It therefore stands to reason that discarding sensitive characteristics would have a positive effect on the kNN classifier as it removes complications, while having a negative effect on the decision tree as it then has significantly fewer attributes to act on.

Changing thresholds



(a) The effects of changing the threshold for the kNN classifier.

(b) The effects of changing the threshold for the decision tree classifier.

Figure 6: Threshold plots for both models.

In the process of looking into altering the thresholds in the algorithms, it was decided that a more efficient method would be to examine the effect of the changing threshold on certain attributes. While it is possible to fumble in the dark, changing the threshold step by step and noting its effects, it is more efficient by far to use these plots to decide on a threshold that would result in the fairest model. As can be seen in Figure 1b, higher thresholds result in higher precision, but lower recall, and vice versa.

As the F1 value is the harmonic mean of precision and recall, it was decided that the fairest possible system, and thus the optimal threshold value, is the value that results in the highest possible F1 value. Figure 6a shows that for the kNN model, the most effective threshold is 0.5, and Figure 6b shows that 0.3 is the best threshold for the decision tree classifier.

It is impossible to ensure that any model is truly completely fair, but after testing and examining, this work has resulted in a model that is fairer than it was at the start, which is promising.